DOCUMENT RESUME

ED 137 336 TM 006 136

AUTHOR Hill, Richard K.
TITLE Conducting Linear Regression Analysis When Observations Have Varying Standard Errors of Measurement.
PUB DATE [Apr 77]
NOTE 8p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Educational Assessment; *Multiple Regression Analysis; *School Districts; *Standard Error of Measurement; State Programs; *Statistical Analysis; *Test Interpretation
IDENTIFIERS California Assessment Program

ABSTRACT

An assumption underlying multiple linear regression is that the standard error of measurement is equal for all observations. The literature has not addressed the procedures to be used when this assumption is violated. It was clear that data analysis to be performed on districts in California would severely violate this assumption, since district mean scores were to be the criterion, and districts vary tremendously in size. The statistical techniques that were developed to conduct these analyses are described. (Author)

Conducting Linear Regression Analysis When Observations Have Varying Standard Errors of Measurement

Richard K. Hill
California Department of Education

Introduction

To facilitate interpretation of test results provided by the California Assessment Program, districts are provided with a statistic called the comparison score band. This band is computed in two steps: first, a predicted score is obtained for the district by regressing a series of input variables collected about the school (called background factors) on the school mean test score, and then the standard error of estimate is added to and subtracted from the predicted score. While such a statistical procedure is routine in most applications, it is not so here.

An assumption underlying multiple linear regression is that the standard error of measurement is equal for all observations. But because districts in California vary greatly in size (Los Angeles City Unified School District tests 43,000 pupils per grade annually, while some rural districts test just one), it is clear that this assumption is grossly violated, and the consequences are extreme and observable. For example, the multiple correlation between the background factors and third grade mean test scores is less than .6 when computed across all districts in California. That same correlation becomes almost .8 when districts testing fewer than 10 pupils--about 10 percent of the districts--are removed from the analysis. Thus, standard regression procedures leave one with two unsatisfactory alternatives: either use all districts, both large and small, in the regression analysis, thereby allowing the large measurement error associated with small districts to mask the relationships that actually exist between the background factors and the criterion, or arbitrarily eliminate smaller districts from the regression process.

Even if the regression problem were to be solved, and a predicted score could be fairly computed for each district, it still would be incorrect to add and subtract the same standard error of estimate for all districts. It has been observed for some time that the mean square residual is far greater for small districts than for large ones. For this reason, a procedure which employed the same standard error of estimate for all districts would greatly underestimate the error for small districts, while greatly overestimating the error for large districts.

A review of the literature revealed that this problem has not been addressed. While literature is replete with examples of regressions done using data from individuals, or using means of groups of equal size, no solutions to this particular problem are published. Forsyth, for example, in his published techniques for conducting such regressions for the state of

Iowa, completely ignored the problem; similarly, meetings with statisticians from departments of education throughout the country revealed that they had a similar awareness of the problem, but also were unsure as to what a correct data analysis would be. This paper describes the techniques that have been developed over the past two years in California to conduct regression analyses and to compute the associated standard error of estimate for each observation.

Computing the Multiple Linear Regression Line

As pointed out in the introduction, an assumption underlying multiple linear regression analysis is that the standard error of measurement is equal across all observations. The violation of this assumption has serious consequences on the regression analyses run at the district level by the California Assessment Program. If the regression were to be computed using all districts in California, the multiple correlation between predictors and the criterion would be around .6. If districts testing fewer than 10 pupils per grade were to be eliminated from that same analysis, the multiple correlation would jump to almost .8. This occurs because of the large amounts of error associated with both the predictors and the criterion in small districts. The large amounts of randon error introduced by these districts into the computation of the linear regression equation obscures the relationship that exists among the predictors and criterion for the vast majority of school districts. Since the results of the regression analysis are reported to all districts, the issue never was whether to do something to make the regressions reflect the actual relationships more accurately; it was a question of what action would be most appropriate.

The Development of a Solution

The first year the analyses were run, the problem was handled simply by eliminating the small districts from the analysis; 106 districts, out of 914 districts throughout California, were eliminated. This was an unsatisfactory solution, however. It seemed unjustifiable for any district, no matter how small, to be completely eliminated. In addition, that solution still gave equal weight to small districts of 20 or 30 pupils per grade and the large city districts. It seemed clear that the most equitable solution would be to compute the regression lines making use of some weighting scheme. The choice of a weighting scheme, however, did not seem to be straightforward.

The first weighting scheme tried was done by weighting all districts by the number of students tested. This procedure resulted in multiple correlations of .99--a value unrealistically high. The result probably occurred because of the great size and deviation from the mean on both predictors and criterion of Los Angeles City.

Current Practice

The search for a realistic approach to conducting the regression analysis concluded after reconsideration of why the problem existed in the first

place. Since it was the differing standard errors of measurement that were at the heart of the problem, it seemed reasonable to use that statistic in the weighting. The final solution, and current practice, is to use the reciprocal of the standard error of the mean as the weighting factor. This method produces a result highly satisfactory on all counts: the size of the multiple correlation is reasonable (around .85), all districts are included, and larger districts can be assured that they had a heavier weight in the determination of the statewide regression line.

Computing the Standard Error of Estimate of the Predicted Score

When a predicted score is generated for a district by the California Assessment Program, a value is added and subtracted from that score to produce a band rather than a point estimate. The band is desired to be of a size such that 25 percent of the districts score below their comparison score band, 50 percent score within, and 25 percent score above.

It had been observed for several years that the size of the band should be dependent on district size. Larger districts have less measurement error in both their predictors and criterion scores and should have smaller bands. If all districts were to receive bands of the same width, most large districts would score within their comparison score band, while few small districts would. At one time, districts were divided into three groups--small, medium and large--and assigned a band width accordingly. While this relatively crude procedure produced acceptable results, an investigation was conducted to see if a more sophisticated and precise way could be established for determining the appropriate standard errors of estimate.

The Development of a Solution

The development of an equation to calculate the standard error for schools of a fixed size required first that a reasonable model of the standard error be posited. The first model tried assumed that the variance error was inversely related to the number of pupils tested in a school; i.e., that $\sigma_E^2 = \frac{\sigma^2}{N}$.

If this model had been correct, then it would have been true that a plot of log σ_E^2 vs. log N would be linear. Such a plot was made by grouping districts of similar size and calculating the variance of the residuals. The relationship was not linear, and the search for an effective model continued.

It was clear that one reason for the failure of the first model was that districts of large size do not have residuals approaching zero. Any good model would have to take into account that there is an asymptotic approach of the residuals to some small but finite value as N increases. This line of reasoning led to the generation of a second model, one which actually was used to report data during the 1973-74 school year. The model posited two variances: the first, called the variance of testing error, was

considered to be inversely related to the number of pupils tested; while the second, called the variance of prediction, was assumed to be constant for all districts. As an equation,

$$\sigma_E^2 = \frac{\sigma_{TE}^2}{N} + \sigma_p^2 \qquad (1)$$

In most applications of linear regression the two error terms would not be examined separately since, in the more typical case, it is reasonable to assume that the variance of measurement error is equal for all observations. There is nothing to be gained by separating the two variances, and they are left combined. In such a case, the variance error of estimate would be calculated as follows:

$$\sigma_E^2 = \sigma^2 (1 - R^2) \qquad (2)$$

However, in this case, it seemed clearly inappropriate to use such a procedure. Since none of the necessary equations for computing the variance error when measurement changes is available in the literature, a stopgap procedure was employed for the reporting of results of the California Assessment Program for the 1973-74 school year. This procedure worked well, and is detailed in the succeeding paragraphs for those who might be interested. A more sophisticated procedure was developed subsequently, and the explanation of that procedure concludes this paper.

The procedure for the 1973-74 school year simply involved computing the median absolute residuals (expressed as a standard score) for all sizes of districts. These median absolute residuals were plotted, as in Figure 1, and a curve was drawn to estimate their values. Then two points were drawn and σ_{TE}^2 and σ_p^2 were solved for.

These results were used as a first approximation. Then, σ_p^2 and σ_{TE}^2 were varied slightly to see if a better fit to the medians could be obtained. These modified values became the parameters of this error-variance equation after being multiplied by the variance of test scores (to correct for the fact that these were standard scores).

For example, the median absolute residuals for second-grade pupils, district by district, were calculated and a line was drawn to fit these points. The following values were generated from the line:

Number of pupils tested in the second grade	Median absolute residual
10	.65
28	.40
50	.36
75	.34
100	.31

Using the values for N = 10 and N = 100, the following two equations were generated:

$$.95063 = \sigma_p^2 + \frac{\sigma_{TE}^2}{10}$$

$$.21623 = \sigma_p^2 + \frac{\sigma_{TE}^2}{100}$$

The solution is $\sigma_{TE}^2 = 8.16$ and $\sigma_p^2 = .135$.

This equation yields the following values:

Number of pupils tested in the second grade	Median absolute residual	$.67\sqrt{.135 + \frac{8.16}{N}}$
10	.65	.65
28	.40	.44
50	.36	.36
75	.34	.33
100	.31	.31

Because the value for N = 28 was thought to vary from .40 by too much, a variety of constants was tried. The best fit seemed to come using $\sigma_{TE}^2 = 7.03$ and $\sigma_p^2 = .146$. This set of parameters yielded the following results:

Number of pupils tested in the second grade	Median absolute residual	$.67\sqrt{.135 + \frac{8.16}{N}}$
10	.65	.61
28	.40	.42
50	.36	.36
75	.34	.33
100	.31	.31

Since these calculations are in standard scores, the estimates of σ_p^2 and σ_{TE}^2 were then multiplied by the variance of mean test scores. The final values for σ_p^2 and σ_{TE}^2 were 16.20 and 779.9, respectively.

The procedure outlined above produced quite satisfactory results. About 50 percent of the large districts and small districts both were scoring within their comparison score band. However, it was desired that this procedure be improved upon for a variety of reasons: it was time-consuming, both to compute the medians and plot them, it was subject to observer bias, and it was, frankly, a very inelegant solution. In addition, such a procedure would not be satisfactory to use in a situation in which there were not a large number of data points. Even with the large number of districts in California (over 900 with second graders), the medians of the subgroups had substantial random error associated with them.

The problem of potential observer bias seemed critical. To draw an analogy, when one observes a scatterplot, it is difficult to draw one regression line. Often, several lines appear as though they could describe the data equally well. As a consequence, the definition of a best-fitting line has been presented and accepted; and computation of a regression line is a straightforward procedure. This problem is very similar. Several lines can be drawn to describe the relationship between district size and the standard errors of estimate. What was needed was some way to compute a value for the standard error directly from the data, without resorting to plots.

Going back to equation 1, it is clear that the mean σ_E^2 can be computed for all districts simply by squaring the residual for each district and dividing by the number of districts. Since σ_p^2 is presumed to be constant for all districts, it can be computed if σ_{TE}^2/N can be computed.

As an estimate to this term, the standard error of the mean ($\sigma_{\bar{x}}^2$) was computed for each district and then the statewide average was calculated ($\overline{\sigma}_{\bar{x}}^2$). Thus, σ_p^2 could be estimated by

$$\hat{\sigma}_p^2 = \overline{\sigma}_E^2 - \overline{\sigma}_{\bar{x}}^2 \qquad (3)$$

From this point, the σ_E^2 for any district could be computed by adding the variance error of measurement to the estimated variance of prediction. While this procedure seemed to be reasonable, it did not work. Although about 50 percent of the districts statewide scored within their comparison score band, fewer than 50 percent of the small districts scored within and more than 50 percent of the large districts did.

Current Practice

The problem seemed to be that the $\overline{\sigma}_{\bar{x}}^2$ was too small. And in fact, it was reasonable that it was too small. Only the measurement error associated with the criterion was being considered; the predictor variables certainly had error associated with their estimates as well (larger error for smaller districts, smaller error for larger districts). It would seem as though a more precise equation for estimating the variance of prediction would be

$$\hat{\sigma}_p^2 = \sigma_E^2 - \Sigma\, \overline{\sigma}_{\bar{x}_i}^2 \qquad (4)$$

where $\Sigma\, \overline{\sigma}_{\bar{x}_i}^2$ is the sum of all the variances of measurement error, both for the criterion and the predictors.

Of course, the straight sum is not appropriate. There is collinearity among the predictors. An approximation of the exact equation is possible by merely considering the standard error of the mean, as in equation 3, but multiplying it by an appropriate constant. Thus,

$$\hat{\sigma}_p^2 = \bar{\sigma}_E^2 - c\,\bar{\sigma}\,\frac{2}{x} \qquad (5)$$

Equation 5 currently is being used by the California Assessment Program in the computation of comparison score bands. The constant is empirically determined, and in different situations varies between 2 and 4.

As a specific example, the results of 1975-76 CAP testing for grade 2 are reported. For that test, the constant used in equation 5 is 2.5. For each district, a predicted score was computed, and then a residual score was computed by subtracting the predicted score from the obtained score. The residual score was then squared. From this value was subtracted the variance error of measurement for that district multiplied by 2.5. That value, called a "difference score," was computed for each district. The mean difference score for the state was 7.5216. Thus, the estimated variance error of prediction for each district was $7.5216 + (2.5 \cdot \sigma\,\frac{2}{x})$. The number of districts scoring above, within or below their comparison score band as a result of the use of this method of calculating the estimated variance error is reported in Table 1.

The largest discrepancies from having 50 percent of the districts scoring within their comparison score band are for the smallest districts (54.9 percent scored within) and the third category (45.5 percent scored within). A chi-square test shows that neither of these values is statistically significantly different at the .05 level from the desired percentage of 50.

Table 1

Number of California School Districts
Scoring Above, Within or Below Their Comparison Score Band
on the Grade 3, 1976, Report of Reading Test Results,
Reported by Size of District

	Size of Districts (Pupils per Grade)				
	1-20	21-50	51-150	151-500	500 +
Above	49 (22.8)*	52 (29.7)	49 (27.5)	45 (25.4)	31 (21.1)
Within	118 (54.9)	83 (47.4)	81 (45.5)	82 (46.3)	74 (50.3)
Below	48 (22.3)	40 (22.9)	48 (27.0)	50 (28.2)	42 (28.6)

* Column percents reported in parentheses